The Power of Hood Friendship for Opportunistic Content Dissemination in Mobile Social Networks

Kanchana Thilakarathna, Aline Carneiro Viana, Aruna Seneviratne, Henrik Petander

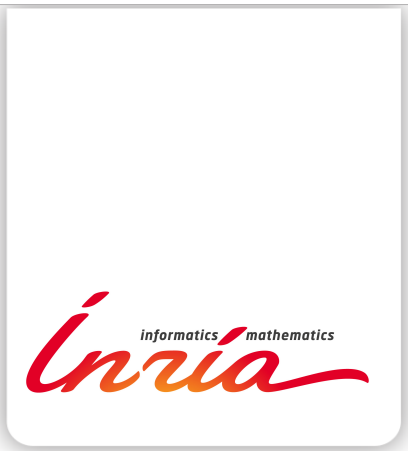
RESEARCH REPORT

N° 8042

August 2012

Project-Teams HIPERCOM

ISSN 0249-6399 ISRN INRIA/RR--8042--FR+ENG



The Power of Hood Friendship for Opportunistic Content Dissemination in Mobile Social Networks

Kanchana Thilakarathna*, Aline Carneiro Viana, Aruna Seneviratne*, Henrik Petander*

Project-Teams HIPERCOM

Research Report n° 8042 — August 2012 — 19 pages

Abstract: We focus on dissemination of content for delay tolerant applications/services, (i.e. content sharing, advertisement propagation, etc.) where users are geographically clustered into communities. Due to emerging security and privacy related issues, majority of users are only willing to share information/content with the users who are previously identified as friends. In this environment, opportunistic communication will not be effective due to the lack of known friends within the communication range. In this paper, we propose a novel architecture that addresses the issues of lack of trust, timeliness of delivery, loss of user control, and privacy-aware distributed mobile social networking by combining the advantages of distributed decentralised storage and opportunistic communications. We formally define a content replication problem in mobile social networks and show that it is computationally hard to solve optimally. Then, we propose a community based greedy heuristic algorithm with novel dynamic centrality metrics to replicate content in well-selected users, to maximise the content dissemination with limited number of replication. Using both real world and synthetic traces, we show that content replication can attain a large coverage gain and reduce the content delivery latency.

Key-words: content dissemination and replication, opportunistic communication, mobile social networking

* NICTA & UNSW, Sydney, Australia.

RESEARCH CENTRE
SACLAY – ÎLE-DE-FRANCE

Bâtiment Alan Turing
1 rue Honoré d'Estienne d'Orves
91120 Palaiseau

Le pouvoir de l'amitié voisine pour la diffusion de contenu opportuniste dans les réseaux sociaux mobiles

Résumé : Nous nous concentrons sur la diffusion de contenu pour les applications tolérantes au retard où les utilisateurs sont géographiquement groupés dans des communautés. En raison des questions liées à la sécurité et confidentialité, la majorité des utilisateurs ne sont prêts à partager des informations que avec les utilisateurs déjà identifiés comme des amis. Dans ce contexte, la communication opportuniste ne sera pas efficace s'il y a un manque d'amis connus dans la portée de communication radio des utilisateurs. Dans cet article, nous proposons une nouvelle architecture qui répond aux questions liées à la confiabilité, à la ponctualité de la livraison, et à la confidentialité de l'utilisateur en combinant les avantages du stockage distribuée et des communications opportunistes. Nous définissons formellement le problème de réplication de contenu dans les réseaux sociaux mobiles et montrons que ce calcul est difficile à résoudre de manière optimale. Ensuite, nous proposons un algorithme glouton basé sur des communautés avec de nouvelles mesures de centralité dynamiques pour répliquer le contenu dans les utilisateurs préalablement bien sélectionnés. Utilisant à la fois des traces réel et synthétiques, nous montrons que la réplication de contenu peut permettre d'atteindre un large gain de couverture et de réduire la latence de livraison de contenu.

Mots-clés : réplication et diffusion de contenu, communication opportuniste, réseau social mobile

1 Introduction

The past few years have seen rapid growth in the use of free social networking applications such as Facebook, Twitter, Google+ and application that support the distribution of user generated content (UGC) such as YouTube and Flickr. Today, more and more UGC is being generated by mobile devices and this will become the dominant generator of content in the near future. The users expect not only to view the content that is provided by the social networking services and UGC distribution sites on their mobile devices, but also to upload the content they create from their mobile devices directly. This will impact the user in two ways. Firstly, it will exacerbate the problem associated with the exponential increase in mobile data traffic that has been widely predicted [5]. Secondly, it will make the problems associated with users privacy and data ownership even more acute, due to currently popular social networking and UGC distribution services being centralised allowing the service provider full control of the user data [4].

There have been numerous proposals for dealing with the ever increasing mobile data traffic by taking advantage of ubiquitous availability of mobile devices and/or access to heterogeneous networks [2, 14, 19, 7]. Majority of these proposals, either exploit short-range communication among users when they meet each other (i.e. *opportunistic communication*) [7, 2], or the availability of *different type of network infrastructures*, such as WLANs [14, 19]. Besides finding alternative ways to transfer the users traffic (i.e. offloading) and reducing the load on the overloaded networks, both solutions help minimising cost, while short-range communication may provide connectivity when there is no access to a network [12]. As a separate body of work, there have been a number of proposals aimed at addressing the issues associated with privacy and the user losing control of their data. This work has lead to the development of *distributed decentralised social networking* architectures [8, 6, 17, 15], where the users are provided control of their data by enabling an individual user or a community of users to host the data.

Despite offering many advantages for mobile users, short-range opportunistic solutions are not adopted for social networking services such as Facebook, or Google+ for two primary reasons. Firstly, there is an inherent reluctance by the users to interact with strangers, due to prevailing security and privacy related issues in social networking. Secondly, because of timeliness of delivery requirements, i.e. when a user wants to share information, none of the couriers of information may be in the vicinity. Although, this has been partially addressed by the use of the temporal and spatial community sub-structures within a set of users of a social network [2], it is is still a significant drawback. In addition, the opportunistic solutions do not address the issues of loss of control of data nor the loss of privacy. On the other hand, offloading through a different type of network infrastructure offers a generic solution for reducing the load on congested networks. It is also directly applicable for supporting the centralized social networking applications and UGC distribution services. Again, they do not address neither the issues of loss of control of data nor the loss of privacy. Finally, distributed decentralised social networking architectures, in contrast, address the issue of loss of control of data and the loss of privacy. However, in a mobile system, they increase mobile data traffic by requiring the replication data on distributed servers [19]. With the current trend of mobile operators moving to *capped* data plans, the distributed architectures would be prohibitively expensive.

In this paper, we propose a new architecture that addresses the issues of lack of trust, timeliness of delivery, loss of user control, and privacy-aware distributed mobile social networking by combining the advantages of distributed decentralised storage and opportunistic communications. Although such approaches are not new, the proposed method of combining the two ideas is novel. In particular, we propose to exploit content replication to bridge the gap between the user and the mobile social networking *friends* who are not in the vicinity of the user, thereby reducing the content delivery latency and increasing the delivery success rate. The devices that

are used to replicate the content are only selected among the users who are previously identified as friends, thereby preserving the privacy of the users and reducing redundant communication cost, energy and storage of devices. Once the replication is completed, the content is propagated through opportunistic direct communication between wirelessly connected friends (i.e. *hood friendship*). However, there is an obvious trade-off between the delivery performance and replication overhead. Thus, the challenge is to replicate efficiently to achieve maximum content delivery performance with limited replication.

In summary, the paper makes the following contributions.

- It formally defines the content replication selection problem in mobile social networks for selecting the optimal set of devices to replicate the shared content with the goal of maximising content propagation with limited replication. This is shown to be a NP-hard problem.
- It presents a community based greedy algorithm for efficient content replication by taking advantage of routine behavioural patterns of mobile users.
- It proposes dynamic centrality metrics to identify most influential users within a community based on a dynamic contact graph composed of opportunistic user encounters.
- It provides an evaluation of the performance of the centrality metrics and the algorithm by extensive trace driven simulations with both real world and synthetic trace data sets. Results show that it is possible to provide delivery rate of 80% with less than 10% replication and approximately 60% of content can be delivered in less than one day latency using the proposed mechanism.

The remainder of the paper is organised as follows: Section 2 presents the related work and followed by the overview of the proposed system and formalisation of the problem of content replication in Section 3. Section 4 presents the dynamic centrality metrics and community based content replication algorithm. Section 5 evaluates the performance of proposed metrics and content replication algorithm. Finally, Section 6 concludes the paper and discusses future work.

2 Related Work

Lee et al. [14] presents a quantitative analysis of offloading 3G data traffic via WLANs using data traces from smartphone users. The results suggests that it would be possible to offload 65% of mobile 3G data traffic to WLANs on demand and 78% of traffic, if the data transfer could be delayed for more than one hour. However, these offloading methods are only effective if there is a centralised server to offload the data when the opportunity arises and thereby does not stand out as an alternative for distributed social networking.

Diaspora [8] can be considered as the only widely used distributed decentralised social networking architecture. It was estimated to have over 1.5 million users in March 2012 indicating a high demand for privacy-aware social networking. In order to host data, Diaspora users need to set up their own server for hosting their content. In previous work [19], we showed that if a mobile device is used to host the content as proposed by Diaspora, either the availability has to be compromised or incur increased communications costs.

Safebook [6] is based on the concept of decentralisation and collaboration among friends and creates a secure social network. Similar to our proposed system, friends are assumed to be cooperative and friends devices are used for replication to increase the availability. MyZone [15] also deploys user profile replicas on the devices of trusted friends to increase the availability of

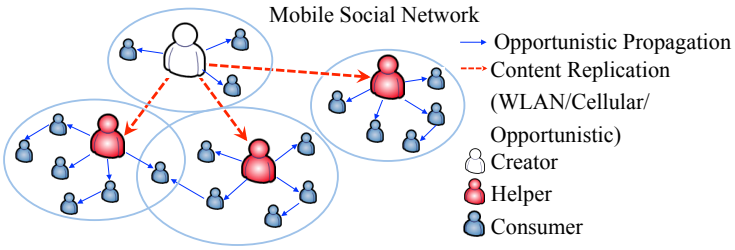


Figure 1: Overview of the system architecture

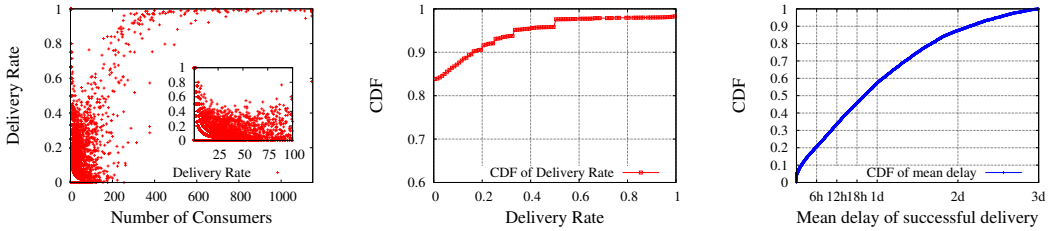
the profile. However, either system are not originally designed for wireless mobile devices where reliability and availability issues are a concern. Further, the availability of content is dependent on the number of replicas. Hence, similar to Diaspora, if used with mobile devices, it would result in increased communication costs and energy consumption of devices. Tribler [17] is a peer-to-peer file sharing system, which takes advantage of social relationships. Again, although there are similarities to our system, Tribler does not consider methods to reduce content delivery latency or minimise communication costs.

Han et al. [7] proposed a target set selection for content replication and there after propagation of the content with opportunistic communication. This mechanism will only work well for popular content. The focus of this work is dissemination of data to/from a centralised data store. Thus, the privacy and censorship issues raised from interaction with strangers are not addressed. Similar to our work, VIP delegation [2] replicates data to a few socially important users in the mobile network. Nevertheless, multi-hop opportunistic propagation is not considered and the metrics used do not consider the dynamic aspects of contact time and duration of users. Further, VIP delegation does not address any privacy and trust related issues in content dissemination.

3 Overview of the System Architecture

The primary objective of our system is to provide privacy-aware distributed mobile social networking which addresses the lack of trust, timeliness of delivery and loss of user control of data. We aim to take advantage of opportunistic direct communication among wirelessly connected friends (i.e. *hood friendship*), and the ever increasing storage and processing power of mobile devices for distributed storage. In this context, our focus is on dissemination of content for delay tolerant social networking applications/services, (i.e. UGC sharing, advertisement propagation, etc.) where users are geographically clustered into communities.

In this section, we provide an overview of our distributed mobile social networking architecture. Suppose a user, namely a *creator*, wants to share a content with a set of users who have previously been identified as *friends* through a social networking service. In a typical distributed content dissemination system, the creator will try to propagate the shared content to the devices of friends. However, each and every friend may not be interested in the shared content even though it is pushed to the device [21]. Assuming that it is possible to identify the probable *consumers* among the friends, we can propagate the content only to the interested consumers and leave the option to other friends to fetch the content from the creator only if they are interested. To this end, our architecture delivers content only to the consumers who are predicted to be interested by the content based on previous consumption patterns. This predictive pre-fetching can be effectively used to mitigate the communication cost and minimise energy and storage of



(a) Content delivery rate (b) CDF of content delivery rate (c) CDF of mean delay

Figure 2: Effectiveness of opportunistic communication in content dissemination in mobile social networks

devices. Even though it is difficult to predict the content demand of a user with a high degree of accuracy, even relatively low levels of accuracy will also improve the cost efficiency of the content delivery compared to implicit propagation to all friends. In addition, we consider that the users are only willing to trust known friends for content propagation. For further cost efficiency, we exploit the opportunistic communication only among the creator and the consumers for content propagation.

Then, the challenge is the timeliness of the delivery due to limited number of consumers in the vicinity of the creator. The idea is to replicate the content to some carefully selected consumers, namely *helpers*, and then trust these helpers to propagate the content to other consumers through opportunistic communication, as shown in Fig. 1. Since helpers are only selected among the consumers, the privacy of the users will be better preserved. Further, this strategy does not consume unnecessary communication cost, storage and energy of devices. As shown in Fig. 1, initial content replication is carried out by either a pre-existent network infrastructure such as WLANs and cellular networks, or short-range opportunistic communication such as WiFi and Bluetooth, if the helpers are in the vicinity of the creator. No matter where the creator and the helper are, the creator has the option of instant replication over whatever network connection available or the replication could be scheduled to be performed through low-cost networks such as WLAN, when it is available. At the same time, the creator initiates content dissemination to the consumers in the vicinity via short-range opportunistic communication.

To perform content dissemination among the creator, the helpers and the consumers, a peer-to-peer (P2P) protocol such as a modified version of BitTorrent, could be used. In particular, a consumer becomes a *propagator* or a seeder only after the consumer has downloaded all pieces of the content. This is to prioritise the energy consumption of mobile devices firstly for its own purposes (i.e. to finish downloading the full content for its own use) and then secondly, to help others in content propagation. The use of such P2P mechanism is outside the scope of this paper.

3.1 Motivation for content replication

In order to investigate the effectiveness of opportunistic communication for content dissemination using only interactions among the creator and the consumers, we first analyse the contact traces generated from Dartmouth campus data sets [13], as described in [3]. This data set was selected because it has been widely used to evaluate many opportunistic communication systems in the literature. We considered two months of data from January to March 2004 which resulted in contact patterns of 1146 users. Further, in order to have more accurate results, we derived a realistic content creation and access model, which is described below.

Each user in the Dartmouth trace data set generates content over two months as indicated

Table 1: Summary of content creation and access model parameters

Content creation model	
Amount per week	142MB [5]
File Size	Gamma (scale=2, mean=4MB) [1]
Inter-arrival time (IAT)	Exponential (mean=3.5 hours) [14]
Content access model	
Content popularity (Number of Consumers)	Pareto Type II (80-10 rule) [21]
Consumer	Random
Transfer rate	2 Mbps [14]
Delivery deadline	3 days

in Table 1. Cisco [5] has predicted that an average smartphone will consume 2.6GB of data per month by 2016. Though the amount of UGC is predicted to increase, the ratio between upload and download has been remained around the 25-75% mark for last few years, according to Cisco [5]. Therefore, we assumed that the average smartphone user generates 142MB per week. The generated file size is characterized by a Gamma distribution, as observed for YouTube [1]. It has been shown that the internet access times of mobile users are exponentially distributed [14]. Therefore, the inter-arrival-time for the content generation is obtained from an exponential distribution with a mean of 3.5 hours such that it spreads throughout the whole trace duration.

The number of consumers is selected based on Pareto distribution since both content popularity in YouTube and degree distribution in Facebook follow power-law models [21]. For this evaluation, it was further assumed that the consumers are randomly distributed and collaborative, i.e. once content is downloaded by a consumer, he is willing to share it with other consumers as described in the previous section.

If content can be opportunistically disseminated from a creator to a consumer through one-hop wireless communication before the delivery deadline of 3 days, we consider it as a successful delivery and otherwise a failure. Fig. 2 shows the simulation results for the Dartmouth trace for content creation and consumption model in Table 1. The main factor which affects the delivery rate is the number of consumers as shown in Fig. 2a. For a large number of consumers, the delivery rate is almost 100% as there are enough consumers to collaborate in content dissemination. However, the number of consumers are often lower in UGC sharing and social networking due to power-law distribution [21], which makes the delivery rate very low for the majority of the instances. As shown in the cumulative distribution function of delivery rate in Fig. 2b, the delivery rate is zero for approximately 84% of content. For each successful delivery, the content delivery delay from the content generation time is illustrated in Fig. 2c. The probability of mean delay being less than one day is approximately 60%.

These results show that the content dissemination using opportunistic communication among the creator and the consumers is not effective when the number of consumers are low, which is the most probable case due to power-low distribution of number of consumers. In such cases, we aim to use content replication to improve the performance of the content delivery. Since content replication to helpers consumes more network resources due to the use of infrastructure based communication, there is an obvious trade-off between delivery performance and content replication overhead. Thus, the content replication challenge is to *maximise the content delivery rate with limited content replication.*

3.2 Problem Statement

Consider a dynamic contact graph $G_t = (C, E_t)$ that changes its topology over time $t \in (1, 2, \cdots, n)$, where t is the minimum duration in which there is no change in the topology. C is a set of consumers and an edge $e \in E_t$ exists among two consumers if they are in the communication range of each other at time t. Suppose a creator u_c wants to share a content via a mobile social networking application/service with consumers $u \in C$. A consumer is called *covered* if it has received the full content before the content delivery deadline of $\Delta > t$ time slots. Let $\alpha^t_{u,v}$ be the effective contact duration between consumers u and v at time t and $\alpha_T(u)$ be the minimum contact duration required by a consumer u to transfer the full content. Consumers are assumed to be collaborative and they become content propagators only after being covered. We denote $P_t \in C$ as the set of content propagators at time t. When there is no initial replication, $P_1 = u_c$. Hence, the set of consumers covered by a creator u_c is:

$$\sigma(u_c) = \left\{ u \in C : \left[\sum_{t=1}^{\Delta} \sum_{v \in P_t} \alpha^t_{u,v} \right] \geq \alpha_T(u) \right\} \tag{1}$$

Consider a set of helpers $H(u_c) \in C$ for a creator u_c. Thus, the creator and the helpers are the initial set of propagators, $P_1 \leftarrow H(u_c) \cup u_c$. The objective is to maximise the number of consumers covered by the creator u_c with a limited number of helpers $\lambda(u_c)$. Then, our CONTENT REPLICATION (CR) problem is to find the maximum cardinality set of $\sigma(u_c)$ with limited number of helpers, formally;

$$\begin{array}{ll} \text{maximise} & |\sigma(u_c)| \\ \text{subject to} & |H(u_c)| \leq \lambda(u_c) \end{array} \tag{2}$$

Here, we show that the CR problem is computationally NP-Hard even for a simple instance of a static social graph.

Theorem 1. CR *is NP-Hard even when* $\Delta = 1$ *and* $\alpha_T(u) = 1$ *for all* $u \in C$.

Proof. Let $G'' = (V'', E'')$ be an undirected graph. A *vertex cover* is a $V_c \subseteq V''$ such that every $(u'', v'') \in E''$ is incident to at least a $u'' \in V_c$. For a given positive integer k, the decision problem of $\exists V_c$ of size at most k is one of the Karp's original 21 NP-Complete problems [10]. We show that vertex cover is polynomial time reducible to CR problem.

Given the vertex cover V_c of size k, we define a specific instance of CR problem as $H(u_c) \leftarrow V_c$, $\lambda(u_c) = k$ and $G_t = G''$. When $\Delta = 1$ and $\alpha_T(u) = 1$, a consumer $u \in C$ is covered if it is connected to at least a $v \in H(u_c)$. This follows that if there is a solution to CR problem, i.e. $\exists\ H(u_c) \leq \lambda(u_c)$ such that every $u \in C$ is incident to at least a $v \in H(u_c)$, $\exists\ V_c$ of size at most k for G'', there is a solution to vertex cover problem. □

4 Content Replication Algorithm

In this section, we present our content replication strategies by taking advantage of routine behavioural patterns among mobile users. Opportunistic encounters among devices are highly dependent on user mobility patterns, which essentially demonstrates social behaviour of users. Hence, there is a diurnal correlation of opportunistic encounters among users. These patterns have been extensively analysed in the areas of context-aware services and mobile social networking [2]. Usually, social behaviour of majority of users have weekly routines. Further, there is high probability that a user meets the same people at the same time in every week. To this end, we aim to take advantage of predictive regularity of encounter patterns of users for the purpose of

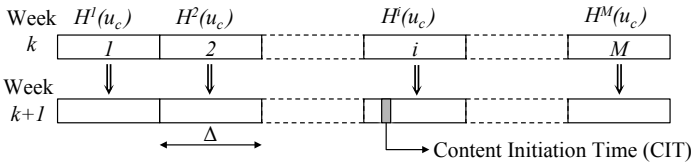


Figure 3: Periodic weekly helper selection

content replication selection. In order to facilitate instant content dissemination, helpers have to be selected in advance. We select helpers for week $k+1$ during the week k as shown in Fig. 3. The week (Δ_w) is divided into Δ time slots, where the Δ is the content delivery deadline. Since we do not know when creators are going to generate content during the week $k+1$, we select several sets of helpers $H_{k+1}(u_c) = \left\{H_k^j(u_c) : u \in C \text{ and } j \in [1 : \Delta_w/\Delta]\right\}$ during the week k. At the end of every week, a central management entity performs helper selection and inform all creators the respective sets of helpers. A creator will be assigned a new set of helpers only if the creator changes its behavioural pattern. In the remainder of this section, we present two helper selection algorithms that can be used for different types of environment and under different resource constraints.

4.1 Greedy helper selection algorithm

Since the problem is NP-Hard, it can not be solved in polynomial time. However, the objective can be approximated efficiently using heuristics. Thus, we present a greedy algorithm GREEDY-HELPERS for CR problem. The most influential user in the network is the one who can cover maximum number of consumers, i.e. has $\max|\sigma(\cdot)|$, which can be intuitively used as a greedy choice property. Algorithm 1 presents the naive greedy algorithm to select $H(u_c)$ for the content creator u_c.

Algorithm 1 GREEDY-HELPERS$(G_t, \Delta, \lambda, u_c)$

1. $D \leftarrow H(u_c) \leftarrow \emptyset$
2. **for** all $u \in C$ **do**
3. Find $\sigma(u)$
4. $D \leftarrow \sigma(u_c)$
5. $H(u_c) \leftarrow u_c$
6. **while** $|H(u_c)| \leq \lambda(u_c)$ or $D \neq C$ **do**
7. Let $u \in (C \setminus (H(u_c) \cup D))$ maximising $|\sigma(u)|$
8. $H(u_c) \leftarrow u$
9. $D \leftarrow D \cup \sigma(u)$
10. **return** $H(u_c)$

D is the set of consumers covered by the creator and the selected helpers. After calculating $\sigma(u)$ for all $u \in C$, i.e. line 3, D will be equaled to the set of consumers covered by the creator $\sigma(u_c)$ (i.e. line 4) and the helper set $H(u_c)$ will be equaled to the creator u_c (i.e. line 5). Then, we loop through until we cover all devices or reach the threshold of replication $\lambda(u_c)$ while selecting the consumer with highest $|\sigma(u)|$ from the remaining consumers. This set-covering flavoured solution has considerably high level of approximation factor. In [11], a similar greedy algorithm is used for *influence maximisation problem* and show that this is $(1-1/e)$ approximation, where e is the base of the natural algorithm. Even though, this provides an acceptable approximation

algorithm for CR problem, finding $\sigma(u)$ for all $u \in C$ is computationally too complex in a dynamic network under resource constraints. To this end, we propose computationally simple dynamic centrality metrics for greedy choice that exploit the temporal and spatial regularity of social wireless connectivity patterns.

4.2 Dynamic Centrality Metrics

As the first step, we aggregate every contact for a single graph without loosing any temporal information. Let an aggregated weighted graph $G = (C, E)$ consists of all edges in G_t, $\forall t \in (1, 2, \cdots, n)$ such that $G = G_1 \cup G_2 \cup \cdots \cup G_n$ and $\alpha^t_{u,v}$ be the edge weights at time t of each $e \in E_t$. For instance, if $\exists\ (u, v, \alpha^1_{u,v} = 20) \in E_1$ and $(u, v, \alpha^2_{u,v} = 30) \in E_2$, there are two edges in E connecting u and v with the contact duration 20 and 30 and happening at $t = 1$ and $t = 2$. Then, we focus on centrality metrics in G having an impression similar to $\sigma(\cdot)$, i.e. expected number of covered consumers.

Hereafter, we propose two kinds of centrality metrics: local and global. Local metrics consider the information available locally, i.e. on-hop away, to decide the influence of the user. The main advantage is its simplity and distributed calculation. Global metrics consider the whole network topology in order to decide the centrality value of the user, which is more complex and needs to be carried out in a central location.

4.2.1 Local metrics

One of the simplest centrality metric that implies the capability of neighbourhood coverage is the degree centrality $C_{LD}(u) = |N(u)|$ where $N(u)$ is the set of neighbours of u in the aggregated graph G. Degree centrality identifies popular nodes in the network and thus has higher influence on content propagation in static networks.

However, simple degree centrality does not guarantee that all counted encounters are practically realisable due to the lack of consideration of temporal information in dynamic networks. Further, the contacts that happen early are important in propagation than those that happen later. Hence, a centrality metric which captures temporal information could be more realistic to be considered in dynamic networks. To this end, we define the initial contact time as $I(u, v) = min\{t\} : \alpha^t_{u,v} > 0$ for all $t \leq \Delta$ and the total contact duration $D(u, v) = \sum_{t=1}^{\Delta} \alpha^t_{u,v}$ for an edge (u, v). We calculate the weight $w_{u,v} = I(u, v) + (1/D(u, v))$ for all $(u, v) \in E$. $w_{u,v}$ has the meaning of earliness and solidity of the contact (u, v). In practice, each mobile device can calculate w locally for all other devices it encounters for a given period. To this end, we define an improved dynamic degree centrality metric:

$$C_{LID}(u) = |N(u)| + \frac{|N(u)|}{\sum_{v \in N(u)} w_{u,v}}$$

$N(u)$ is the set of neighbours of u. C_{LID} has the impression of how early and how independently the user makes other users into content propagators. We aim to use C_{LID} in the greedy choice for CR problem.

4.2.2 Global metrics

Even though centralised systems have disadvantages in terms of privacy and scalability, we make use of the global information to perform more accurate heuristics. Here, we define two path-based centrality metrics for node ranking. We first define a naive simple metric $C_{GP}(u) = \sum_{v \in C} p(u, v)$ for all $t \leq \Delta$ where $p(u, v) = 1$ if there is a path between u and v and $p(u, v) = 0$ otherwise. This can be viewed as an extended degree for node u giving heuristics about the popularity and

the reachability of the node. Simplicity of the metric is the main advantage, which requires only information about existence of a path.

On the other hand, simplicity does not provide accurate heuristics. Therefore, we define a more complex dynamic centrality metric by considering temporal information such as the contact duration $D(u, v)$ and the initial contact $I(u, v)$. We construct a directed aggregated graph $G' = (C', E')$ by directing all edges in G for both directions with same weights. Next, we prune all unrealisable edges in the network, i.e. if a content is to be propagated via a node, its outgoing contact has to take place after at least its first incoming contact. At this point, we have an aggregated graph and at each node there is a guarantee that content will be propagated to other nodes if the content has arrived at the node. We calculate shortest-path $sp(u, v)$ for all node pairs $(u, v) \in C'$ in terms of edge weights $w_{u,v} = I(u, v) + (1/D(u, v))$. We define a path-based dynamic centrality metric C_{GIP}, similar to C_{LID}, such that it implies how early and how independently the user makes other content propagators as:

$$C_{GIP}(u) = \sum_{v \in C'} p(u, v) + \frac{\sum_{v \in C'} p(u, v)}{\sum_{v \in C'} sp(u, v)}$$

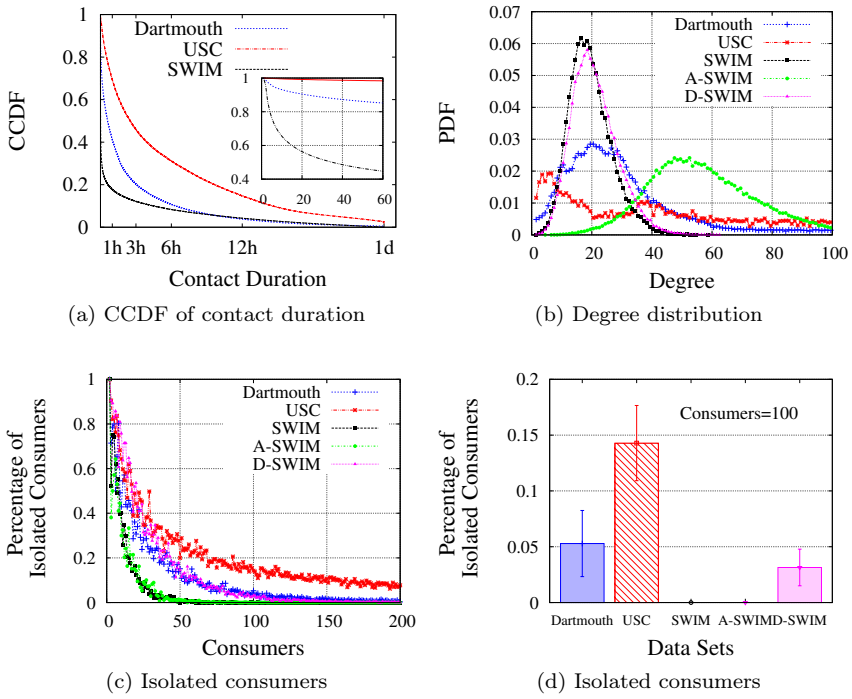
4.3 Community based greedy algorithm

In this section, we present our content replication algorithm which combines social sub-structural properties such as communities with previously defined dynamic centrality metrics.

Algorithm 2 COMMUNITY-GREEDY$(G, G', \Delta, \lambda, u_c)$

1. $D \leftarrow H(u_c) \leftarrow \emptyset$
2. **for** all $u \in C$ **do**
3. Find centrality metric $C_{LD}(u), C_{LID}(u), C_{GP}(u), C_{GIP}(u)$
4. communities $\leftarrow$ k-clique-algorithm$(G', 3)$
5. Let a community $com(u)$ be the u's community
6. $D \leftarrow com(u_c)$
7. $H(u_c) \leftarrow u_c$
8. **while** $|H(u_c)| \leq \lambda(u_c)$ or $D \neq C$ **do**
9. Let $u \in (C \setminus (D \cup H(u_c)))$ maximising $C_{LD}(u), C_{LID}(u), C_{GP}(u), C_{GIP}(u)$
10. $H(u_c) \leftarrow u$
11. $D \leftarrow D \cup com(u)$
12. **return** $H(u_c)$

In order to distribute helpers within the network, we extract social sub-structures present in the contact graph. For this we detect communities using *k-clique* community algorithm. Then, we distribute helpers among communities based on their ranking given by the proposed dynamic centrality metrics as in Algorithm 2. First, the consumer with highest centrality value is selected as a helper and rely on that helper to propagate the content with in the community. Then, the next highest consumer from a different community is selected, i.e. line 9 of the Algorithm 2. If the threshold of replication is lower than the number of communities, initial content propagators will not be selected from the creator's community, assuming that the creator is capable to propagate the content within its community. There can be a scenario where the majority of the consumers do not belong to communities. Then, the selection is purely based on the centrality value of consumers. In the next section, we evaluate the performance of this approach with respect to different centrality metrics.



(a) CCDF of contact duration

(b) Degree distribution

(c) Isolated consumers

(d) Isolated consumers

Figure 4: Dynamics of trace data sets

5 Performance Evaluation

First, we present the dynamics of mobility traces and simulation setup that we use for performance evaluation. Then, we evaluate the benefits of the content replication and show that it can incentivise users to use mobile social networks for UGC sharing. Finally, we compare the performance of different centrality metrics for selecting influential consumers in terms of delivery success rate and latency.

5.1 Mobility Trace Data Sets and Simulation Setup

In order to evaluate the proposed strategies, we use publicly available data sets, that contain wireless connectivity patterns of users that are in the communication range within each other.

5.1.1 Experimental data sets

Two real-world data sets are considered: Dartmouth data set [13] described in Section 3.1 and USC [9] data set, which also contains wireless connectivity patterns of users in a campus environment with 1846 users in average per week. Moreover, we use the synthetic traces that are generated using the SWIM simulator [16] by extending the Cambridge campus data set [18]. SWIM is the 500-nodes extended version of the original Cambridge data set of an experiment with 36 Bluetooth enabled iMotes. Then, the trace data is further scaled up to generate 1500-nodes versions: 1) D-SWIM by keeping the density constant and 2) A-SWIM by keeping the area constant. We use these extended versions to understand the performance variation of the

content replication in different environmental conditions.

For both Dartmouth and USC, we consider that two users are in contact when they are connected to the same WiFi access point as described in [3]. For SWIM, when two users are within the Bluetooth communication range of each other, we consider those two devices are in contact. As per the complementary cumulative distribution function of contact duration in Fig. 4a, more than 50% of users have more than 3 hours of contact duration per day. This can be considered as very high value which can be used to transfer any mount of data between two users, if the delivery latency is one day. Even in Dartmouth more than 80% of users have more than 60 seconds contact duration per day. In contrast, SWIM has much lower contact duration. The degree distribution of all data sets is shown in Fig. 4b. SWIM and A-SWIM have highly skewed degree distributions, i.e. majority of the users have the similar number of contacts. In contrast, USC has a fairly distributed degree distribution while the degree distribution of the Dartmouth lies in between those two extremes. Further, we analysed the amount of isolated consumers when we randomly select a set of consumers. Fig. 4c illustrates that USC trace contains a large number of isolated users compared to others. For the case of 100 consumers (Fig. 4d), nearly 15 of them are isolated in USC and only 5 of them are isolated in Dartmouth. In contrast, in all trials there are no isolated consumers in SWIM, it is less than 5 even in A-SWIM. The five trace data sets that we use for this evaluation is not similar and cover various aspects in properties that mainly affect the performance of content replication. Hence, the performance evaluation resulted from these trace data sets would be applicable to a wide variety of social environments.

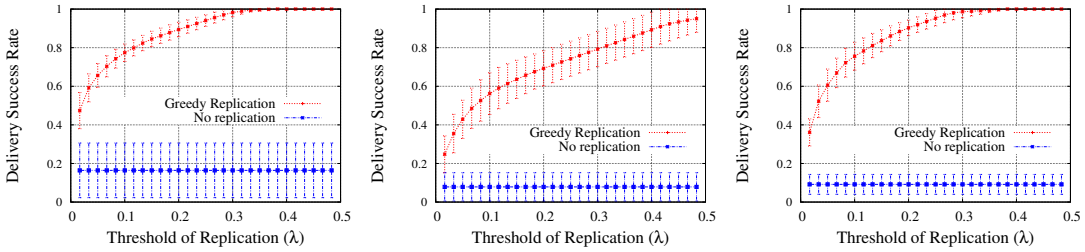
To this end, the five trace data sets that we use for this evaluation is not similar and cover various aspects in properties that are mainly affect the performance of content replication. Hence, the performance evaluation resulted from these trace data sets would be applicable to wide variety of social environments.

5.1.2 Simulation Setup

To exploit weekly routine patterns of users, the trace data sets are divided into weeks. We select the set of helpers according to proposed algorithms during week k, namely *monitoring period*, and evaluate the performance in terms of delivery success rate and delivery latency during week $k+1$, namely *evaluation period*, as illustrated in Fig. 3. The content delivery deadline Δ is considered as 3 days and the creator and consumers are selected randomly. In content propagation, we only use opportunistic contacts among the consumers. Based on conclusions of Section 3.1, we consider that the number of consumers for a creator is 100. Each creator will generate a content of size 8.4MB, which is the median content size in YouTube [1] and the transfer rate among consumers are considered as uniform and 2Mbps [14]. Hence, a consumer has to have aggregated contact duration $\alpha_T(u)$ of 33.6 seconds with the creator or any of the helpers or the propagators to completely download the content. All simulations are carried out varying the monitoring and evaluation periods through out the duration of the data sets.

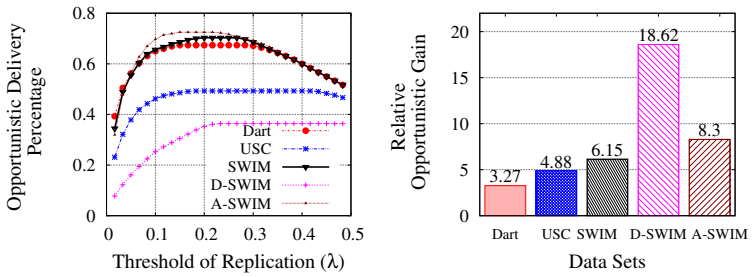
5.2 Benefits of Content Replication

In this section, we evaluate and compare the content delivery rate and delivery latency given by content replication against the case where no replication is performed. Moreover, the increment in the percentage of one-hop opportunistic transmission is also analysed, which is proportional to energy and communication cost savings.



(a) Dartmouth Data Set (b) USC Data Set (c) SWIM-500 Data Set

Figure 5: Delivery success rate against the threshold of replication (λ)



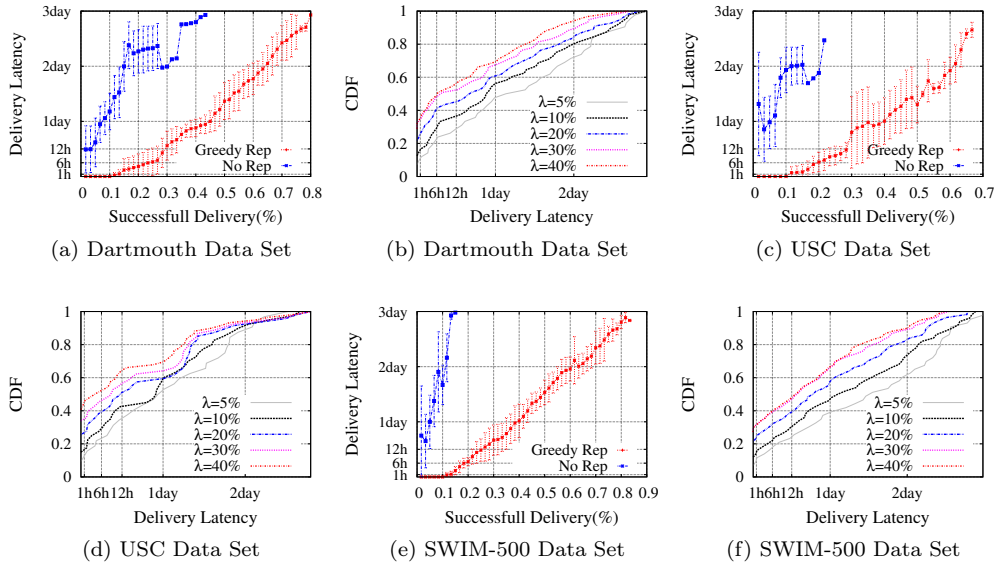
(a) (b)

Figure 6: Analysis of amount of opportunistic content delivery. $\lambda = 10\%$

5.2.1 Delivery Success Rate

Fig. 5 compares the delivery success rate for the GREEDY-HELPER (Algorithm 1) and for no replication approaches. For all data sets, it is evident that there is a significant gain in content delivery compared to no replication approach. In Dartmouth (Fig. 5a), the delivery success rate is approximately 80% for 10% of content replication, while USC has a comparatively lower success rate of approximately 60%. Note that after a certain level of replication (i.e. approximately $\lambda = 0.1$), the delivery rate shows linear increment, i.e. further replication will not deliver content to any other consumer via opportunistic communication because there will be only isolated consumers to be selected for content replication. Since the main goal of content replication is to increase the content delivery via opportunistic communication, this threshold of replication can be considered as an effective upper bound for λ.

We also analyse the amount of successful deliveries via opportunistic communication. This is shown in Fig. 6a. It shows that the percentage of opportunistic deliveries increases with λ, only for low λ values. In Dartmouth, it is possible to deliver content for approximately 70% of the consumers via opportunistic communication compared with below 20% when there is no replication. We extended our simulations to the two extended versions of the SWIM data set to understand the behaviour of opportunistic delivery percentage. The results are closely related to the degree distribution of the data sets as shown in Fig. 4b. D-SWIM has the lowest performance because it was extended by increasing the area and number of users while keeping the density of the network constant and equal to SWIM. This makes the network more spread and increases the number of appearing communities and consequently requires a high level of replication to cover the same number of consumers as in SWIM. In contrast, when we increase the density as in



(a) Dartmouth Data Set (b) Dartmouth Data Set (c) USC Data Set

(d) USC Data Set (e) SWIM-500 Data Set (f) SWIM-500 Data Set

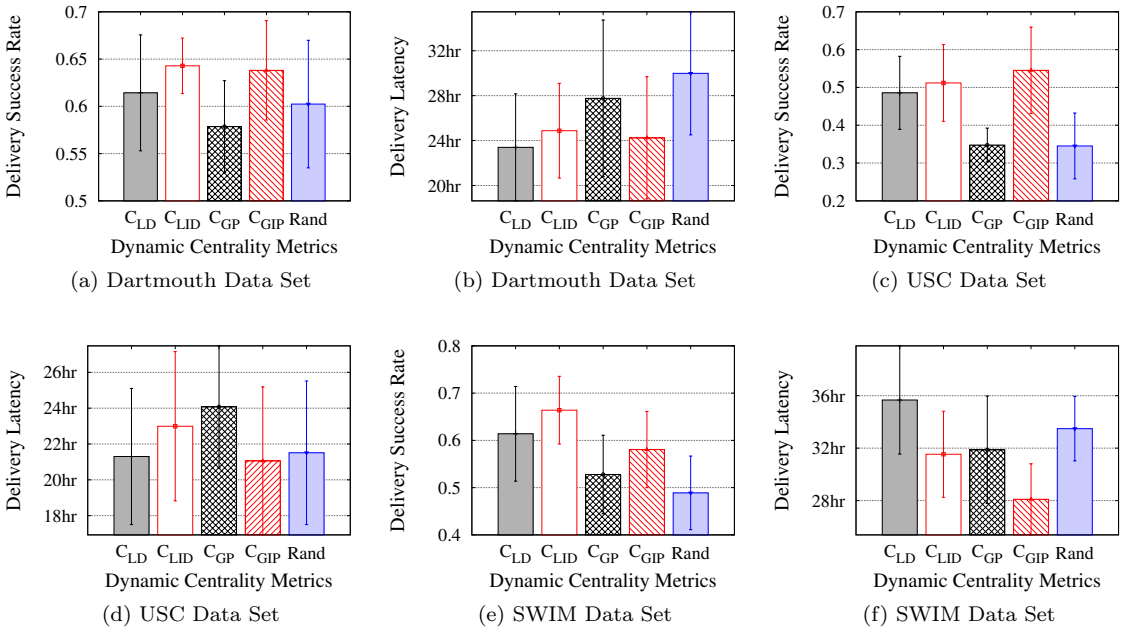
Figure 7: Delivery latency for percentage of successful delivery. $\lambda = 10\%$

A-SWIM, it improves the opportunistic delivery percentage. Similarly, in USC, there is a large number of isolated users as shown in Fig. 4d, which decreases the overall density compared to the other data sets. Hence, the density of the contact graph has a considerable impact on the opportunistic delivery performance.

Fig. 6b summarises the relative gain in opportunistic communication with respect to the no replication approach for all data sets. Even though D-SWIM has the lowest percentage of opportunistic delivery, it has the highest relative gain of 18.62. This happens because in D-SWIM, the percentage of opportunistic delivery when there is no replication is as low as 1.3%. Dartmouth data set shows the lowest gain of 3.27 times, since it consists of well connected users compared to other considered environments. Thus, the results show that it is possible to significantly increase the one-hop opportunistic delivery success rate among consumers with a low number of initial content replication, i.e. approximately less than 10% of the consumers. On the other hand, increment in opportunistic transmission is proportional to the energy and communication cost and thereby incentivise mobile users to take part in distributed mobile social networks for content dissemination.

5.2.2 Delivery Latency

Even though we are dealing with delay tolerant content dissemination applications, delivery latency is still one of the prime factor to consider in mobile social networking. Fig. 7 shows the time taken by greedy replication and no replication approaches against the percentage of successful delivery. In Dartmouth (Fig. 7a), content replication delivers content to 40% of the consumers in less than 1 day, whereas the latency is almost 3 days if there is no initial content replication. All three data sets show similar behaviour in terms of delivery latency. For instance, the time taken to cover 40% and 60% of the consumers is approximately 1 day and 2 days respectively, in all three cases. The cumulative distribution function of the delivery latency for successful deliveries are shown in Fig. 7b, 7d and 7f. In all data sets, the probability of the



(a) Dartmouth Data Set (b) Dartmouth Data Set (c) USC Data Set

(d) USC Data Set (e) SWIM Data Set (f) SWIM Data Set

Figure 8: Comparison of different dynamic centrality metrics. $\lambda = 10\%$

delivery latency being less than 1 day is approximately 60% for $\lambda = 10\%$. In [20], it has been observed that 55% of Flicker content is uploaded after a lag of more than 1 day. Thus, we believe that the delivery latency resulted from the opportunistic approach is practical in such content dissemination applications.

For applications/services that require a lower delay, it is possible to increase the threshold of content replication as shown in Fig. 7. The common pattern is that the delivery latency reduces with increasing λ values. In Dartmouth and USC, there is a 20% increment in the delivery latency being less than 1 day when we increase λ from 5% to 30% while it is a 30% increment in SWIM. However, the difference between two consecutive λ values becomes smaller when λ increases. For instance, the cumulative distribution functions of delivery latency for $\lambda = 30\%$ and $\lambda = 40\%$ are almost identical for SWIM and Dartmouth. Again, similar to the delivery success rate, the explicit replication will not increase the delivery latency as much, i.e. there is an upper bound for λ which does not increase the delivery latency significantly after that. Thus, in these particular mobile social environments, the content replication shows promise to significantly increase the delivery success rate and reduce the delivery latency compared to the no replication approach. Since our trace data sets contains a diverse contact patterns among a substantially large number of users in three different cities and time periods, we believe the proposed content replication strategy would hold for in a wide variety of environments.

5.3 Comparison of Dynamic Centrality Metrics

In this section, we compare the influence of different dynamic centrality metrics in content replication selection that are defined in the Section 4.2. Further, we compare the results with *Random* selection of helpers, which is the simplest way of selecting helpers without any knowledge about the contact patterns among consumers.

Fig. 8 shows the content delivery success rate and the mean content delivery latency when the helpers are selected based on different centrality metrics according to the Algorithm 2. When we compare the two local centrality metrics, C_{LID} has a slightly better delivery success rate with lower standard deviation than C_{LD} in all three data sets. Similarly, the improved global centrality metric C_{GIP} has better performance in terms of both delivery rate and latency compared to the naive C_{GP}. This is due to the fact that each improved metrics, C_{LID} and C_{GIP}, consider the time dependency in connectivity patterns, which affects the content propagation. However, there is no significant difference between the performance of local and global metrics in general. On the other hand, although the random selection has a bit worse performance than the improved metrics, random selection is not negligible due to the heavy reduction in resource requirements.

All these general similarities are related to dynamics of the contact patterns among consumers in these environments. For instance, when the degree of the majority of the consumers are similar, i.e. skewed degree distribution as shown in Fig. 4b for SWIM, the performance of a simple degree based local centrality metric becomes significant as depicted in Fig. 8e for the same data set. In contrast, when the degree distribution is not skewed, the intelligent path based selection will perform better, similar to C_{GIP} performance for Dartmouth and USC. In particular, USC has the highest increment of approximately 20% in coverage for C_{GIP} compared to random because USC has the most distributed degree distribution. On the other hand, due to the large number of isolated consumers (Fig. 4c), USC does not have much gain in delivery latency. In SWIM, C_{GIP} has much lower delivery latency compared to random selection because it has the lowest number of isolated users.

Therefore, the selection of the appropriate centrality metric to identify the most influential users in content dissemination is highly environment dependent. In some cases, the random selection even without any knowledge about contact patterns would perform better or similar to intelligent selection. Han et al. [7] also observed similar behaviour in their target set selection for content dissemination, where the random selection performs similar to the greedy solution.

6 Conclusion

In this paper, we proposed a novel opportunistic content dissemination architecture that addresses the issues of lack of trust, timeliness of delivery, loss of user control, and privacy-aware distributed mobile social networking by using content replication. We formally defined the content replication problem in mobile social networks and show that this is NP-hard. Starting from this, we developed a community based greedy algorithm for efficient content replication by taking advantage of routine behavioural patterns of mobile users. Using both real world and synthetic traces, we showed that content replication can attain a delivery success rate of 80% with less than 10% replication and approximately 60% of the content can be delivered in less than one day latency. Thus, the proposed strategies can be used to incentivise mobile users to take part in distributed mobile social networks for content dissemination. Further, different dynamic centrality metrics were proposed to identify most influential users within a community and to show that the performance are highly environment dependent.

In future work, we aim to analyse the content replication load distribution among helpers such that highly influential helpers may not need compromise their resources to help others. Further, a social-aware content propagation among consumers will be exploited again for the purpose of communication cost and energy cost reductions.

References

[1] A. Abhari and M. Soraya. Workload generation for youtube. *Multimedia Tools and Applications*, 46(1):91–118, 2010.

[2] M. Barbera, J. Stefa, A. Viana, M. de Amorim, and M. Boc. Vip delegation: Enabling vips to offload data in wireless social mobile networks. In *DCOSS'11*, pages 1–8. IEEE, 2011.

[3] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of human mobility on opportunistic forwarding algorithms. *Mobile Computing, IEEE Transactions on*, 6(6):606–620, 2007.

[4] G. Chen and F. Rahman. Analyzing privacy designs of mobile social networking applications. In *Embedded and Ubiquitous Computing, 2008. EUC'08. IEEE/IFIP International Conference on*, volume 2, pages 83–88. IEEE, 2008.

[5] Cisco. Cisco visual networking index: Global mobile data traffic forecast update, 2011–2016. In *http://www.cisco.com*.

[6] L. Cutillo, R. Molva, and T. Strufe. Safebook: A privacy-preserving online social network leveraging on real-life trust. *Communications Magazine, IEEE*, 47(12):94–101, 2009.

[7] B. Han, P. Hui, V. Kumar, M. Marathe, J. Shao, and A. Srinivasan. Mobile data offloading through opportunistic communications and social participation. *Mobile Computing, IEEE Transactions on*, (99):1–1, 2011.

[8] http://joindiaspora.org.

[9] W. jen Hsu and A. Helmy. CRAWDAD data set usc/mobilib (v. 2008-07-24), July 2008.

[10] R. M. Karp. Reducibility among combinatorial problems. In *Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972.

[11] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[12] A. Khan, V. Subbaraju, A. Misra, and S. Seshan. Mitigating the true cost of advertisement-supported âĂIJfreeâĂİ mobile applications. 2012.

[13] D. Kotz, T. Henderson, I. Abyzov, and J. Yeo. CRAWDAD data set dartmouth/campus (v. 2009-09-09). Downloaded from http://crawdad.cs.dartmouth.edu/dartmouth/campus, Sept. 2009.

[14] K. Lee, I. Rhee, J. Lee, S. Chong, and Y. Yi. Mobile data offloading: how much can wifi deliver? In *Proc. of the Co-NEXT '10*, pages 1–12, Philadelphia, 2010.

[15] A. Mahdian, J. Black, R. Han, and S. Mishra. Myzone: A next-generation online social network. In *Tech Report: Department of Computer Science, University of Colorado at Boulder*, 2011.

[16] A. Mei and J. Stefa. Swim: A simple model to generate small mobile worlds. In *INFOCOM 2009, IEEE*, pages 2106–2113. IEEE, 2009.

[17] J. Pouwelse, P. Garbacki, J. Wang, A. Bakker, J. Yang, A. Iosup, D. Epema, M. Reinders, M. Van Steen, and H. Sips. Tribler: a social-based peer-to-peer system. *Concurrency and Computation: Practice and Experience*, 20(2):127–138, 2008.

[18] J. Scott, R. Gass, J. Crowcroft, P. Hui, C. Diot, and A. Chaintreau. CRAWDAD data set cambridge/haggle (v. 2009-05-29), May 2009.

[19] K. Thilakarathna, H. Petander, J. Mestre, and A. Seneviratne. Enabling distributed social networking on smartphones. In *Accepted to Proc. of ACM MSWiM*, oct 2012.

[20] I. Trestian, S. Ranjan, A. Kuzmanovic, and A. Nucci. Taming user-generated-content in mobile networks via drop zones. In *INFOCOM'11*, pages 2840–2848, 2011.

[21] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow. The anatomy of the facebook social graph. *Arxiv preprint arXiv:1111.4503*, 2011.

Contents

1 Introduction **3**

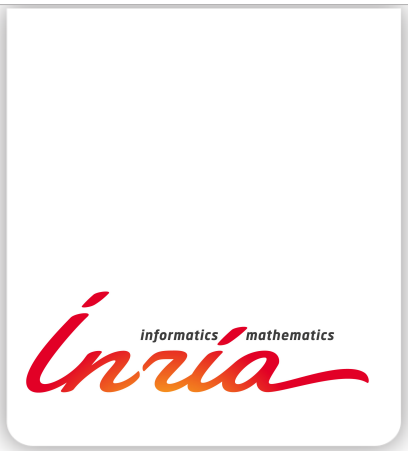
2 Related Work **4**

3 Overview of the System Architecture **5**
3.1 Motivation for content replication 6
3.2 Problem Statement 8

4 Content Replication Algorithm **8**
4.1 Greedy helper selection algorithm 9
4.2 Dynamic Centrality Metrics 10
4.2.1 Local metrics 10
4.2.2 Global metrics 10
4.3 Community based greedy algorithm 11

5 Performance Evaluation **12**
5.1 Mobility Trace Data Sets and Simulation Setup 12
5.1.1 Experimental data sets 12
5.1.2 Simulation Setup 13
5.2 Benefits of Content Replication 13
5.2.1 Delivery Success Rate 14
5.2.2 Delivery Latency 15
5.3 Comparison of Dynamic Centrality Metrics 16

6 Conclusion **17**



RESEARCH CENTRE
SACLAY – ÎLE-DE-FRANCE

Bâtiment Alan Turing
1 rue Honoré d'Estienne d'Orves
91120 Palaiseau

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399